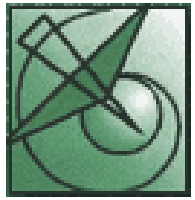
# Odborná terminologie knihovnictví a informační vědy očima uživatelů databáze TDKIV

Předběžné výsledky projektu

Helena Kučerová

VOŠIS Praha



# Cíl projektu:

Využít metody **kvantitativní analýzy**
k objevení nových poznatků a námětů,
skrytých v uchovávaných **uživatelských datech**
o vyhledávání v TDKIV.

- Zadavatel:
  Redakční rada TDKIV
- Řešitel:
  VOŠIS Praha – 6členný studentský tým
- Termín: zimní semestr 2011/2012

# Česká terminologická databáze knihovnictví a informační vědy

- Producent:
  grant MK ČR 2001 – 2002
  NK ČR – tým 50 autorů, lektorů a redaktorů
- Obsah: knihovnictví, knihověda, informační věda, informační technologie

# TDKIV v číslech:

rok vzniku: **2003**

počet (preferovaných) termínů: **3 000**

počet ekvivalentů (nepreferovaných termínů): **4 000**

počet dotazů / rok: **53 000**

počet vyhledávaných výrazů / rok: **12 000**

počet uživatelů (UIP adres): **500**

průměrná doba strávená v databázi: **1 – 2 minuty**

# Struktura terminologického hesla

- **termín**
- anglický ekvivalent
- **ekvivalent** (nepreferovaný termín)
- **příbuzný termín**
- výklad termínu (autorský / normativní)
- zdroj výkladu
- poznámka
- třídník (věcná kategorie)
- autor / konzultant / lektor / redaktor

# Ukázka terminologického hesla

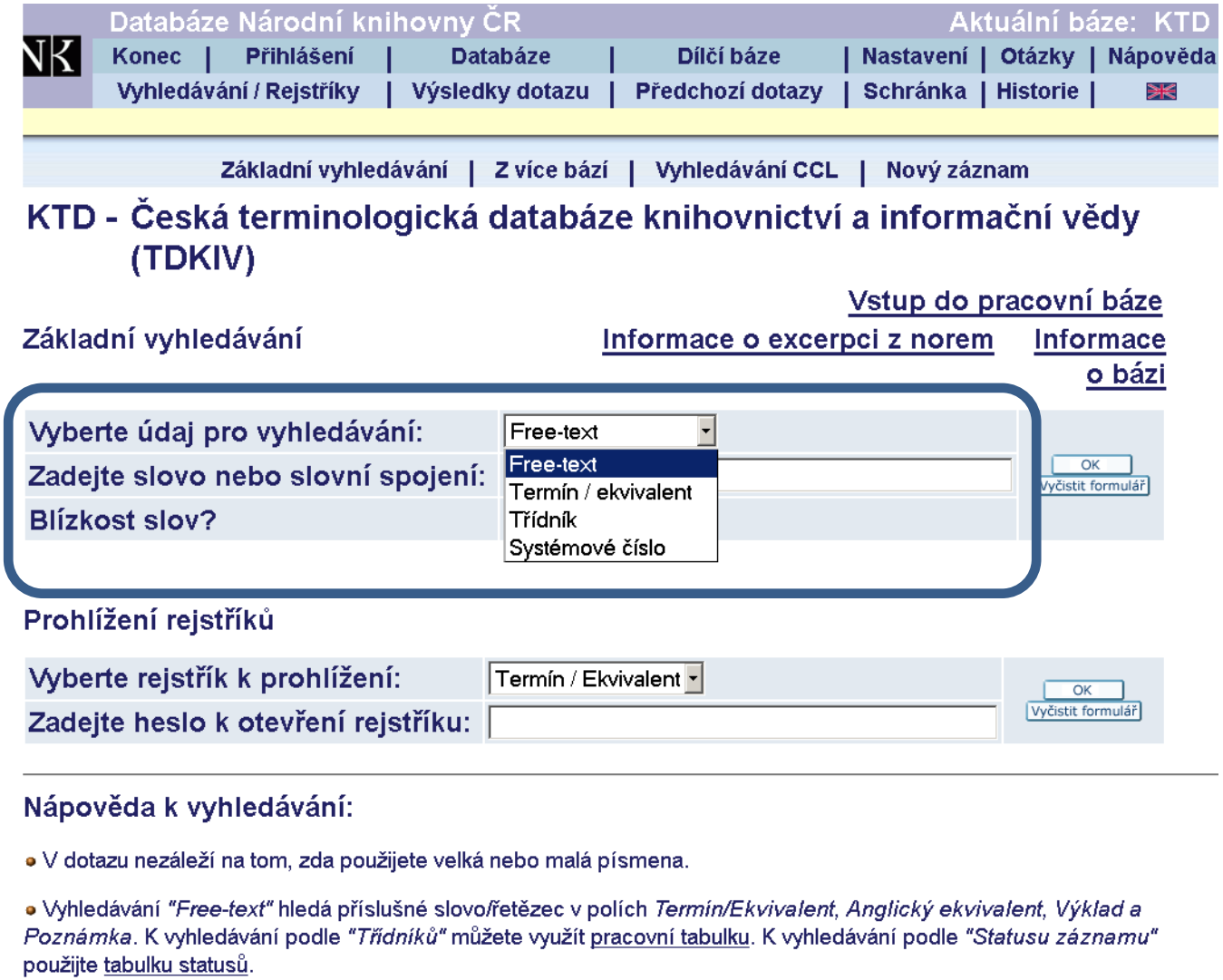
| | |
|---|---|
| **Termín** | •data mining |
| **Poznámka** | ciz., neol., český ekvivalent není ustálený |
| **Ekvivalent** | **dolování dat** (neol.) |
| | text mining |
| | vytěžování informací |
| | vytěžování médií |
| | vytěžování obsahu databáze |
| | web mining (ciz.) |
| **Termín anglicky** | data mining |
| **Výklad termínu** | Technologie vyhledávání, modelování a prezentace předem neznámých informací, příp. znalostí a vztahů mezi daty v rozsáhlých databázích a datových skladech. Analýzy se odvozují přímo z obsahu **dat**, nikoliv na základě hypotéz či dotazů uživatele. Využívají se techniky umělé inteligence (neuronové sítě, rozpoznávání, samoučící se algoritmy), jež mohou být kombinovány s technikami statistického a matematického modelování (klasifikační pravidla nebo stromy, regrese, shluková analýza) a s nástroji OLAP (on-line analytické zpracování).<br>[KUPKA-2002:57-60]<br>[POKORNÝ-1999:96-99] |
| **Zdroj výkladu** | KUPKA-2002:57-60 |
| | POKORNÝ-1999:96-99 |
| **Příbuzný termín** | •datový sklad |
| | •extrakce informací |
| | •knowledge discovery in databases |
| **Viz též** | •záznam v bázi KSL |
| | |
| **Autor hesla** | Kučerová, Helena |
| **Lektor hesla** | Kimlička, Štefan |
| **Redaktor hesla** | Schwarz, Josef |
| | Burgetová, Jarmila |
| | |
| **Systém. číslo** | 000000088 |

# Možnosti vyhledávání v TDKIV

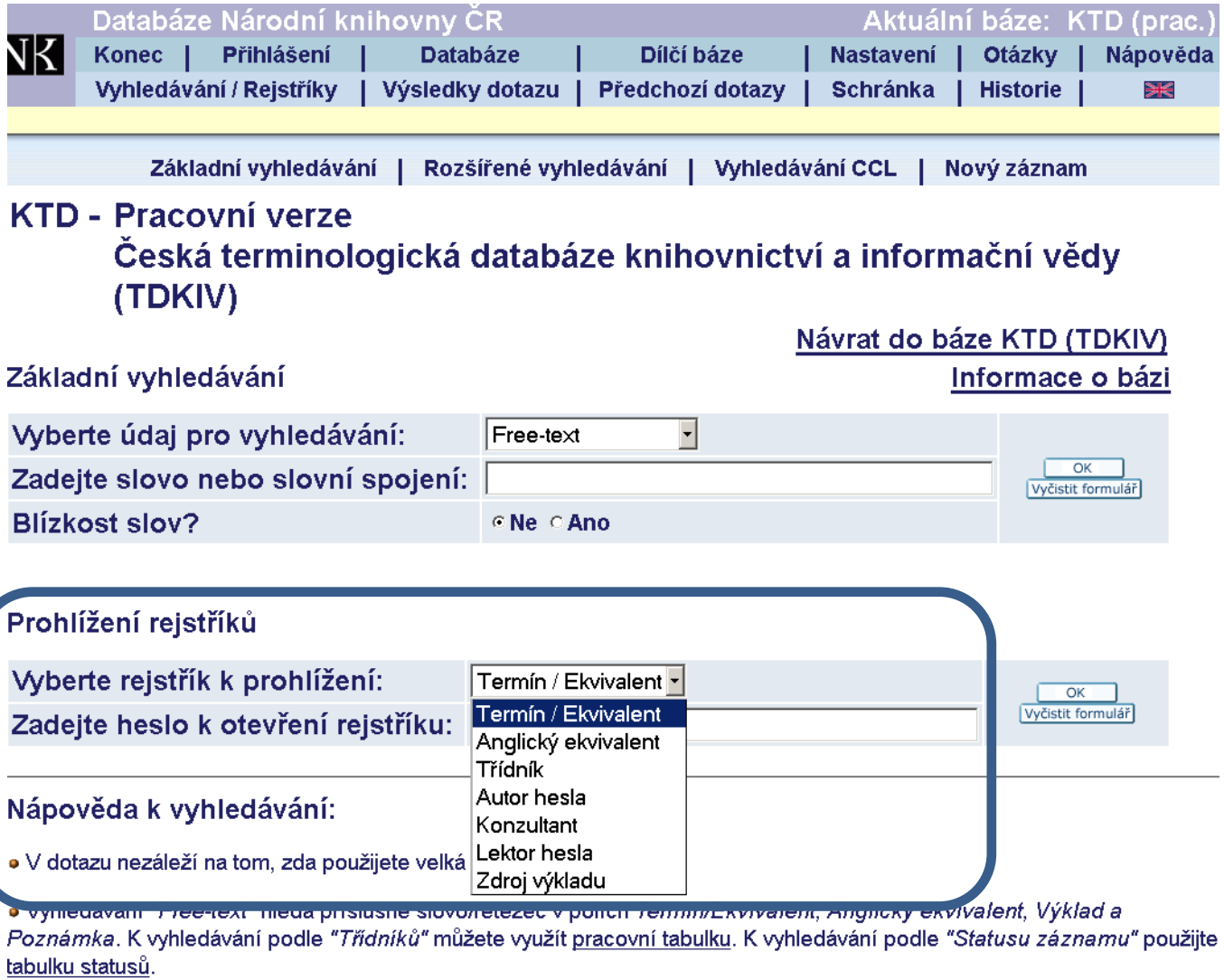
| Způsob vyhledávání | Použitý index | Zdrojová báze |
|---|---|---|
| vyhledávání<br>SEARCH | z více polí<br>FREE TEXT | KTD |
| listování<br>SCAN | z jednoho pole | KTDP<br>Pracovní báze |
| navigace<br>Příbuzný termín | | |

# Vyhledávání v TDKIV

SEARCH



# Listování v TDKIV




SCAN

# Navigace v TDKIV

| | |
|---|---|
| **Termín** | •**řízený slovník** |
| **Termín anglicky** | controlled vocabulary |
| **Výklad termínu** | **Slovník** lexikálních jednotek selekčního jazyka uspořádaný specifickým způsobem (např. zahrnuje vztahy ekvivalence, hierarchie a asociace), který slouží pro indexaci a vyhledávání dokumentů. [KATUŠČÁK,MATTHAEIDESOVÁ,NOVÁKOVÁ-1998:306] |
| **Normativní výklad** | Seznam slov nebo frází, které byly schváleny pro indexování. [ČSN ISO 5127-2003] |
| **Zdroj výkladu** | KATUŠČÁK,MATTHAEIDESOVÁ,NOVÁKOVÁ-1998:306 |
| **Zdroj norm.výkladu** | ČSN ISO 5127-2003 |
| **Příbuzný termín** | •kvalifikovaný Dublin Core |
| | •nepreferovaný termín |
| | •předmětový heslář |
| | •syntetická metoda |
| | •tezaurus |
| | |
| **Autor hesla** | Balíková, Marie |
| **Lektor hesla** | Hrazdil, Aleš |
| **Redaktor hesla** | Schwarz, Josef |
| | Burgetová, Jarmila |
| | |
| **Systém. číslo** | 000001624 |

# Výzkumné otázky:

**Statistický rozbor a vizualizace:**

- četnost hledaných a nalezených / nenalezených výrazů
- četnost způsobů vyhledávání
- čas vyhledávání
- rozdělení IP adres

**Data mining a business intelligence:**

- příčiny neúspěšných dotazů
- společně hledané termíny

# Metodika:

- korpusová terminografie
- kvantitativní analýzy, pokus o data mining (bibliomining)

- logy z vyhledávání v systému Aleph – data za období **březen 2010 – srpen 2011**
  cca **80.000** záznamů

- MS SQL, dotazování v SQL

| | | | | | |
|---|---|---|---|---|---|
| 2010030106062306 | 217.197 | 1 | 21 | KTD | Free-text= sigla |
| 2010030108372010 | 195.113 | 6 | 21 | KTD | Free-text= informační management |
| 2010030108413077 | 147.229 | 0 | 21 | KTD | Free-text= avízo |
| 2010030108423950 | 195.113 | 2 | 21 | KTD | Free-text= ontologie |
| 2010030108425062 | 195.113 | 2 | 21 | KTD | Free-text= ontologie |
| 2010030108425941 | 195.113 | 6 | 21 | KTD | Free-text= informační management |
| 2010030108431765 | 195.113 | 6 | 21 | KTD | Free-text= informační management |
| 2010030108491647 | 195.113 | 1 | 21 | KTD | Free-text= certifikační autority |
| 2010030108501060 | 195.113 | 1 | 21 | KTD | Free-text= certifikační autority |
| 2010030108510436 | 195.113 | 1 | 21 | KTD | Free-text= certifikační autority |
| 2010030108511091 | 195.113 | 1 | 21 | KTD | Free-text= rss |
| | | | | | eriálové publikace |
| | | | | | ntologie |
| | | | | | olandr |
| | | | | | olandr |
| | | | | | orma |
| | | | | | OAN PERIOD |
| | | | | | OAN |
| | | | | | apír |
| | | | | | ýroba papíru |
| | | | | | apírenství |
| | | | | | apírenství |
| | | | | | apírenství |
| | | | | | oan period |
| | | | | | ace |
| | | | | | ourdni |
| | | | | | nihtisk |
| | | | | | nižnično-informačné služby |
| | | | | | konomie model IS-LM |
| | | | | | nižnično-informačné služby |
| | | | | | ermín= ekonomie model IS-LM |
| 2010030109594718 | 158.195 | 0 | 21 | KTD | Free-text= knižnično-informačné služby |
| 2010030110010636 | 158.195 | 25 | 21 | KTD | Free-text= informační služby |
| 2010030110014915 | 158.195 | 25 | 21 | KTD | Free-text= informační služby |
| 2010030110015666 | 158.195 | 0 | 21 | KTD | Free-text= knižniční služby |
| 2010030110025765 | 158.195 | 0 | 21 | KTD | Free-text= knižniční služby |
| 2010030110035919 | 158.195 | 0 | 21 | KTD | Free-text= knižniční služby |
| 2010030110110749 | 010.001 | 1 | 21 | KTD | Free-text= certifikáty |
| 2010030110355599 | 195.113 | 6 | 21 | KTD | Free-text= kognitivní |

```
SELECT a.search_text as hledany_a,
COUNT(a.search_text) as pocet_hledani,
b.search_text as hledany_b
FROM SEARCH_KTD a
JOIN SEARCH_KTD b on a.cas = b.cas
AND a.ip = b.ip and a.datum = b.datum
WHERE a.search_text <> b.search_text
AND a.HITS <> 0 and b.HITS <> 0
GROUP BY a.search_TEXT, b.SEARCH_TEXT,
a.SEARCH_TEXT+a.DATUM+a.CAS,
b.SEARCH_TEXT+b.DATUM+b.CAS
HAVING COUNT(a.SEARCH_TEXT) > 5
```

# Struktura dat

čas zadání dotazu

část IP adresy počítače,
ze kterého byl dotaz zadán

počet nalezených záznamů

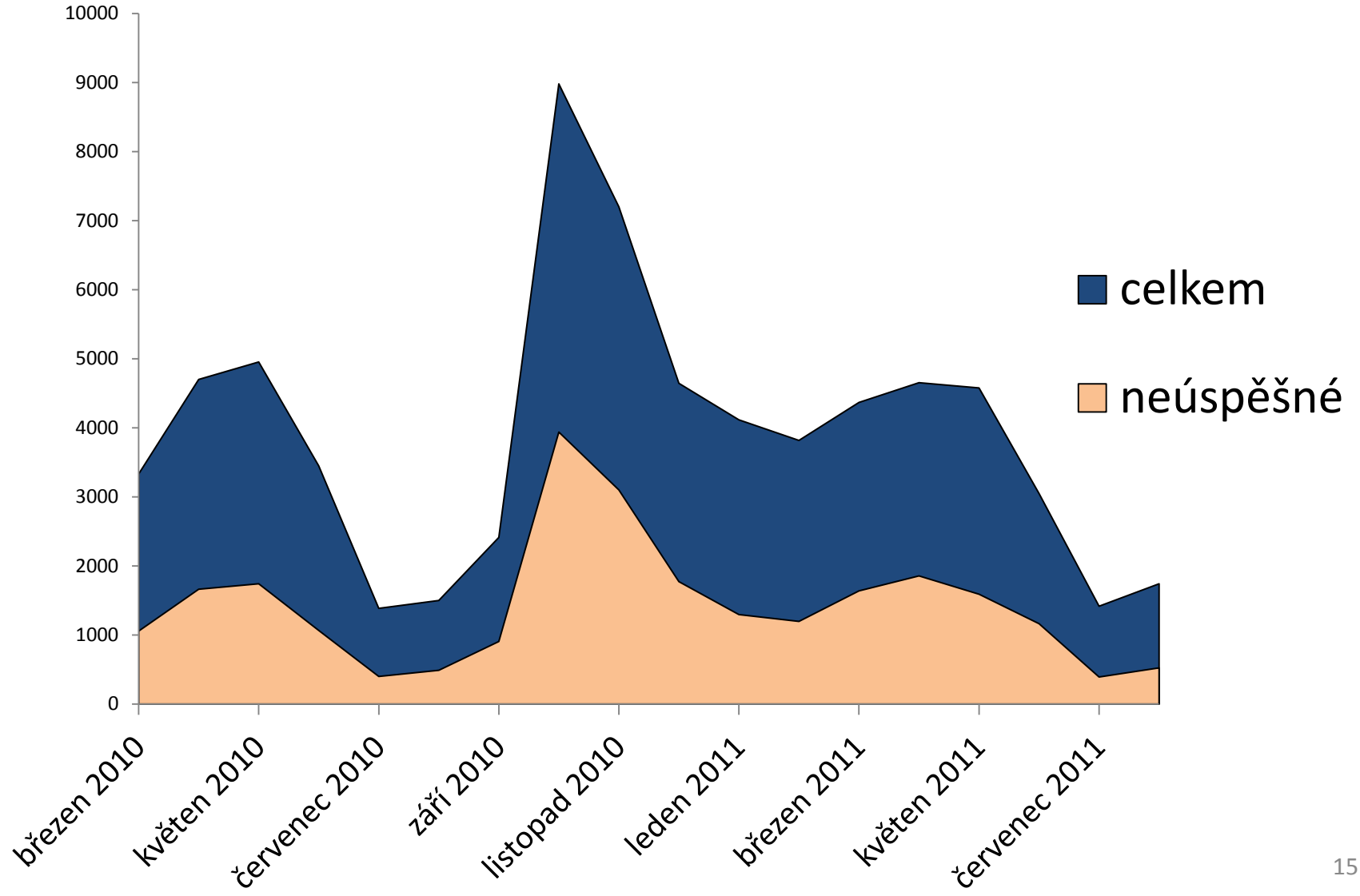
způsob vyhledávání
(21 – základní vyhledávání,
23 – pokročilé vyhledávání,
29 – vyhledávání v rejstřících…)

prohledávaná báze
(KTD, KTDP)

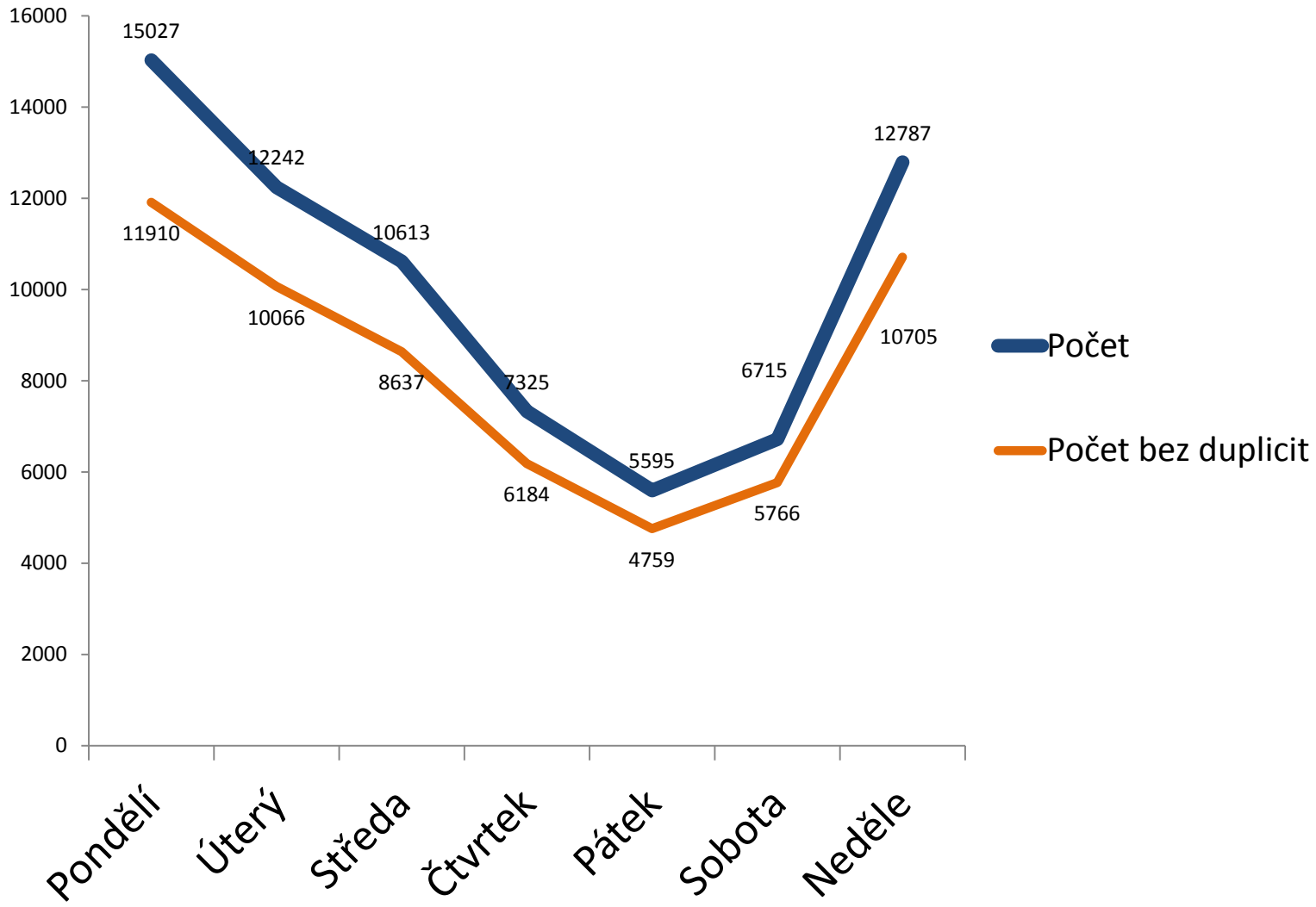
vyhledávaný výraz

| TIME_STAMP | IP | HITS | TYP | SEARCH | SEARCH_TEXT |
|---|---|---|---|---|---|
| 2010030106062306 | 217.197 | 1 | 21 | KTD | Free-text= sigla |
| 2010030108372010 | 195.113 | 6 | 21 | KTD | Free-text= informační management |
| 2010030108413077 | 147.229 | 0 | 21 | KTD | Free-text= avizo |
| 2010030108423950 | 195.113 | 2 | 21 | KTD | Free-text= ontologie |
| 2010030108425062 | 195.113 | 2 | 21 | KTD | Free-text= ontologie |
| 2010030108425941 | 195.113 | 6 | 21 | KTD | Free-text= informační management |
| [illegible] | 195.113 | 6 | 21 | KTD | Free-text= informační management |
| 2010030108491647 | 195.113 | 1 | 21 | KTD | Free-text= certifikační autority |
| 2010030108501060 | 195.113 | 1 | 21 | KTD | Free-text= certifikační autority |
| 2010030108510436 | 195.113 | 1 | 21 | KTD | Free-text= certifikační autority |
| 2010030108511091 | 195.113 | 1 | 21 | KTD | Free-text= rss |
| 2010030109014741 | 147.229 | 9 | 21 | KTD | Free-text= seriálové publikace |
| [illegible] | [illegible] | 2 | 21 | KTD | Free-text= ontologie |
| 2010030109093525 | 193.084 | 0 | 21 | KTD | Free-text= holandr |
| 2010030109100751 | 193.084 | 0 | 21 | KTD | Free-text= holandr |
| 2010030109145152 | 195.113 | 6 | 21 | KTD | Free-text= norma |
| 2010030109151268 | 195.113 | 1 | 21 | KTD | Free-text= LOAN PERIOD |
| 2010030109170096 | 195.113 | 25 | 21 | KTD | Free-text= LOAN |
| 2010030109191401 | 193.084 | 73 | 21 | KTD | Free-text= papír |
| [illegible] | [illegible] | [illegible] | 21 | KTD | Free-text= výroba papíru |
| 2010030109195423 | 193.084 | 0 | 21 | KTD | Free-text= papírenství |
| 2010030109200318 | 193.084 | 0 | 21 | KTD | Free-text= papírenství |
| 2010030109201871 | 193.084 | 0 | 21 | KTD | Free-text= papírenství |
| 2010030109202824 | 195.113 | 1 | 21 | KTD | Free-text= loan period |
| [illegible] | [illegible] | [illegible] | [illegible] | KTD | Termín= anotace |
| 2010030109352534 | 193.084 | 0 | 21 | KTD | Free-text= fourdni |
| 2010030109400995 | 193.084 | 2 | 21 | KTD | Free-text= knihtisk |
| 2010030109574758 | 158.195 | 0 | 21 | KTD | Free-text= knižnično-informačné služby |
| 2010030109584346 | 146.102 | 0 | 21 | KTD | Free-text= ekonomie model IS-LM |
| 2010030109584714 | 158.195 | 0 | 21 | KTD | Free-text= knižnično-informačné služby |
| [illegible] | 146.102 | 0 | 21 | KTD | Termín= ekonomie model IS-LM |
| 2010030109594718 | 158.195 | 0 | 21 | KTD | Free-text= knižnično-informačné služby |
| 2010030110010636 | 158.195 | 25 | 21 | KTD | Free-text= informační služby |
| 2010030110014915 | 158.195 | 25 | 21 | KTD | Free-text= informační služby |
| 2010030110015666 | 158.195 | 0 | 21 | KTD | Free-text= knižniční služby |

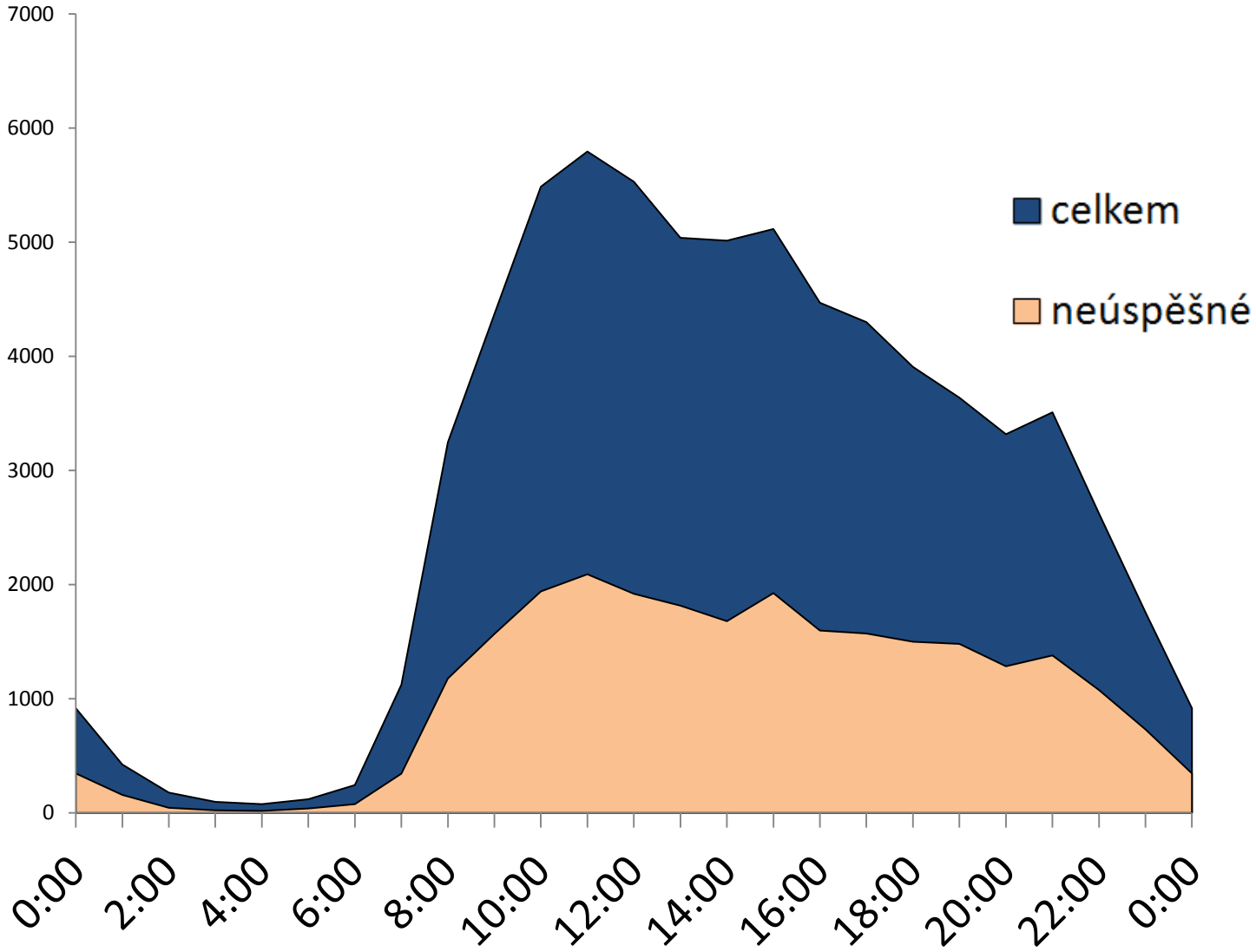
# Vyhledávání podle měsíců



# Vyhledávání podle dní



# Čas vyhledávání



# Rozdělení uživatelů

Vyhledávání v KTD:
42 701 vyhledávání (sessions)
4 264 (cca 10 %) vyhledávání z IP 195.113.xx (PASNET)

| | počet vyhledávání | | počet IP adres | |
|---|---|---|---|---|
| SCAN | 6 026<br>80 % | 80 % | 50<br>21 % | 20 % |
| SEARCH | 28 134<br>80 % | | 112<br>24 % | |
| SCAN | 1 507<br>20 % | 20 % | 181<br>80 % | 80 % |
| SEARCH | 7 034<br>20 % | | 361<br>76 % | |

# Nejvyhledávanější výrazy

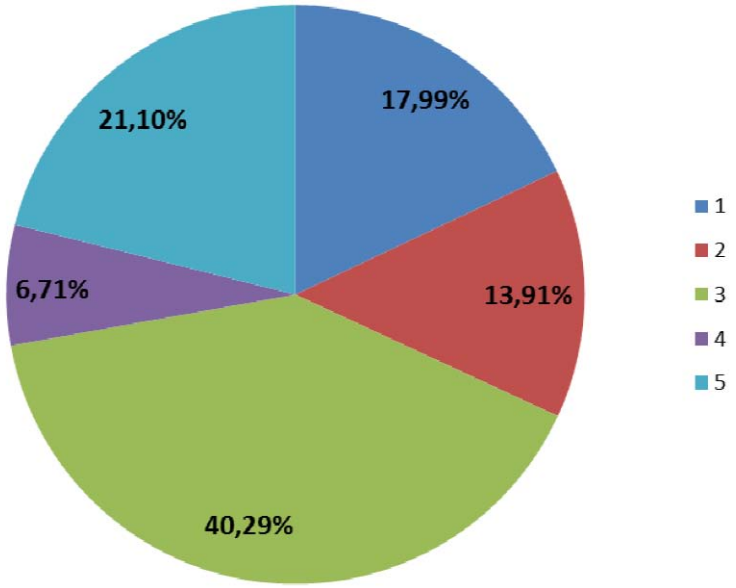
| | |
|---|---|
| Free-text= akvizice | 432 |
| Free-text= Rešerše | 389 |
| Free-text= informace | 370 |
| Free-text= knihovna | 348 |
| Free-text= bibliografie | 264 |
| Free-text= dokument | 228 |
| Free-text= anotace | 225 |
| Free-text= Databáze | 200 |
| Free-text= abstrakt | 199 |
| Free-text= tezaurus | 195 |
| Free-text= Informační věda | 194 |
| Free-text= informační gramotnost | 180 |
| Termín= informace | 177 |
| Free-text= kniha | 164 |
| Free-text= informační zdroj | 159 |
| Free-text= signatura | 155 |
| Free-text= monografie | 154 |
| Free-text= Informační služby | 150 |
| Free-text= katalog | 141 |
| Termín= dokument | 133 |

# Úspěšnost vyhledávání

| počet vyhledávaných výrazů | 17 285 |
|---|---|
| úspěch (>=1 hitů) | **6 867** (cca 40 %) |
| neúspěch (0 hitů) | **10 326** (cca 60 %) |

# Nenalezené výrazy

| | |
|---|---|
| Free-text= informační potřeba | 104 |
| Free-text= AACR2 | 59 |
| Free-text= signet | 57 |
| Free-text= informačný zdroj | 49 |
| Free-text= čtenářský fond | 39 |
| Free-text= sociální síť | 39 |
| Free-text= informační architektura | 38 |
| Free-text= registre | 37 |
| Free-text= pagina | 37 |
| Termín= Registre | 36 |
| Free-text= sfx | 33 |
| Termín= bibliografia | 32 |
| Free-text= jednotná informační brána | 30 |
| Free-text= pozvánka | 30 |
| Free-text= taxonomie | 30 |
| Free-text= facebook | 28 |
| Free-text= elzevirský formát | 28 |
| Free-text= inkunábula | 27 |
| Free-text= česká národní bibliografie | 26 |
| Free-text= repozitář | 26 |




1 překlep
2 doporučeno zařadit do TDKIV
3 věcně nerelevantní k TDKIV
4 nesprávný typ vyhledávání
5 jiný jazyk než čeština

# Výrazy hledané společně

| | | |
|---|---|---|
| pořádání informací | informační analýza | 143 |
| primární dokument | dokument | 91 |
| popisný údaj | bibliografický popis | 70 |
| referát | anotace | 70 |
| věcné zpracování | indexace | 66 |
| katalogizátor | katalogizační záznam | 55 |
| abstrakt | anotace | 54 |
| informační gramotnost | databáze | 48 |
| zpracování dokumentů | dokument | 42 |
| deskriptor | tezaurus | 40 |
| synopse | referát | 30 |
| informační analýza | pořádání informací | 26 |
| jmenné zpracování | věcné pořádání | 24 |

# Předběžné shrnutí výsledků

**Potvrzeno: Kvantitativní metody mohou poskytnout jak náměty ke zkvalitnění použitelnosti a přístupnosti databáze, tak i přímé podněty k terminologické práci**

- náměty na zařazení nových termínů nebo ekvivalentů do TDKIV
- klastry současně hledaných termínů umožňují nacházet sémantické vztahy
- odhaleny problémy k řešení:
  - 2/3 vyhledávání končí neúspěchem
  - 1/5 z nenalezených výrazů jsou překlepy
  - významný podíl dotazů ve slovenštině
  - dotazy na zkratky

# A co na to studenti…

*Rozšířili jsme si znalosti programu MS SQL, získali větší praxi s prací s databázemi, získali jsme informace o dataminingu – co to je, jak se používá, naučili jsme se, že díky dataminingu můžeme z databáze, kterou máme k dispozici získat takové informace, o kterých jsme předtím ani nevěděli, že je můžeme zjistit.*

*Co jsou lidé vůbec schopni napsat do vyhledávače Národní knihovny,*

*Zjistil jsem, že se datamining už v budoucnosti nechci zabývat, není to obor pro mě.*

*Tento projekt mi změnil názor na to, co všechno se dá s daty dělat. Taky to, že z takového malého množství sloupců (kategorií) se dá vyvodit mnoho informací užitečných, které nejdou na první pohled vidět.*

# Plány na další pokračování výzkumu:

- tvorba vícečetných klastrů
- ontologizace TDKIV (syntagmatické → paradigmatické vztahy)
- porovnání vyhledávacích výrazů s termíny v databázi
- analýza vyhledávacích procesů
- „opravdový“ data mining ☺

**Další náměty?**